BUILDING A SAMPLING FRAME FROM MULTIPLE LISTS: IDENTIFYING ORGANIZATIONS IN THE VOLUNTARY AND NOT-FOR-PROFIT SECTOR IN CANADA

Linda Lefebvre & Peter G. Wright [1]

ABSTRACT

Voluntary and not-for-profit organizations are increasingly being recognized for the important role they play in Canadian society, yet efforts to improve the capacity of these organizations to fulfill their missions are constrained by the lack of information about their characteristics. As part of an initiative by the Canadian government to study the not-for-profit and voluntary sector, Statistics Canada conducted, on behalf of the Canadian Center for Philanthropy, a survey to collect baseline information on the function, size and roles of organizations in this sector and the benefits that they work to provide. In order to create the sampling frame for this survey, many list files of not-for-profit businesses and charities from federal and provincial sources had to be put together. Merging of the files and, more specifically, hierarchical exact matching helped to eliminate multiple cases as well as to incorporate additional information to many records on the sampling frame. Through examples, this document presents the practical aspects of file preparation, record linkage and hierarchical matching.

KEY WORDS: Record linkage, hierarchical exact matching

[1] Linda Lefebvre, R.-H.Coats Building, 17th floor, Ottawa (Ontario), Canada, K1A O6T, linda.lefebvre@statcan.ca.
Peter G. Wright, R.-H.Coats Building, 17th floor, Ottawa (Ontario), Canada, K1A O6T, peter.wright@statcan.ca.

1. INTRODUCTION

The objective of the National Survey of Nonprofit and Voluntary Organizations (NSNVO) was to describe the nonprofit and voluntary sector by providing statistical information on size, financial situation and scope of the nonprofit and voluntary organizations in Canada, as well as on their capacity to accomplish their mission and their needs and challenges.

Statistics Canada and its client, the Canadian Center for Philanthropy (CCP), defined the survey's target population in terms of pre-existing lists of registered charities and nonprofit organizations. Charities were defined as registered organizations that are exempt from certain tax rules and can issue tax receipts to Canadian donors. Incorporated nonprofit organizations were identified by specific federal or provincial tax forms they are required to complete or with a special non-profit variable on Statistics Canada's Business Register (BR). A list of voluntary organizations was not readily available but it was considered that the nonprofit organizations and charities would include most organizations that recruited volunteers. CCP acknowledged that the scope of our study would exclude the smaller so-called "grassroots" organizations. It was also determined that the appropriate respondents for the survey (i.e. the sampling units) would be the people working in the local branches of the bigger organizations, not in the head offices.

Data files of registered charities and incorporated nonprofit organizations were provided by Canada Customs and Revenue Agency (CCRA as it was then), provincial tax authorities and Statistics Canada. An approximate count for each of the list files provided is shown in Table 1:

Table 1 – Number of records on each source file provided to create the sampling frame for the NSNVO

SOURCE FILES	Number of records
Registered charities and nonprofit organizations in Canada (CCRA)	82, 000
Provincially registered nonprofit organizations	170, 000
Statistics Canada's Business Register (for stratification information only, such as revenue and main activity)	2, 000, 000[2]

[2] The number of records on the Business Register includes all type of businesses, not only the nonprofit organizations and the charities.

2. CHALLENGES TO ADDING BR INFORMATION TO CCRA RECORDS

Merging of the different source files was necessary to create the sampling frame, to obtain additional information as well as to eliminate possible duplicates. Due to the large number of

records, manual merging was impossible so an automated method using computer programming was used. When files are required to be merged, the ideal solution is to have a common and unique identifier on all source files, preferably a numerical one. This was not always the case and even when such an identifier existed, the level at which it was available may not have been the one needed.

The CCRA lists were the first set of files to be merged with the BR list to obtain additional information. A unique and common identifier existed on both sets of lists, but not at the level of the sampling unit (the local branch). There actually was a unit level identifier at the sampling unit level on each set of files, but it was not a common one (Charity # for CCRA files and Location # on BR file as shown in Table 2 below). Luckily, for tax purposes in Canada, the Business Number (BN) is assigned to each incorporated business and appeared on both files. If a business or organization has no branches, then there is a one-to-one correspondence between the CCRA and the BR files and the BN was used to do a perfect match merge. A challenge arose in situations where an organization would have many branches (called locations) under the same head office (called the enterprise). There was then a many-to-many correspondence between the two files and the BN alone could not be used to do the merging. An example of this is illustrated in Table 2.

Table 2 – Examples of branch identification information on CCRA and BR records that can be used for merging

CCRA records			
BN	Charity #	NAME	ADDRESS
123	0001	AAA	123 MAIN
123	0002	AAA	44 RIVER
555	0001	BOB	BOX 77

BR records			
BN	Location #	NAME	ADDRESS
123	S111	AAA	44 RIVER ST.
123	S222	LLL	GEN. DEL.
123	S333	AAA Inc	5-123 MAIN
888	S531	PGW	6 GIL ST.

In those situations, another approach had to be used. The solution found was to use the name of the organization and/or the address to attempt merging the corresponding locations of the same enterprise from the different source files. But then again, another challenge appeared due to variations in spelling in the name and/or the address of the organizations on the different files. A method was needed to standardize these fields to facilitate the merging process.

In order to standardize the name of an organization, a concatenating function to remove articles like “THE” and “LE/LA”, company types like “LIMITED” and “INC.”, and much more, was created. For example, a name like “LITTLE BROTHER LTD” would become “LITTBROR”.

In order to standardize the address, an in-house generalized program at Statistics Canada called PCODE was used. PCODE concatenates an address into a string, with all parts of the address written backward in a standard order, without spaces and using certain rules of formatting. For example, both “123 YONGE ST., SUITE 500, TORONTO, ONTARIO” and “500-123 YONGE, TORONTO, ONTARIO” would map to “CONTORONTOYONGEST123500”. For some misspelled municipality and street names, PCODE could provide the correct information if the postal code were supplied as input. However some incomplete and misspelled addresses could not be parsed at all.

3. ADDING RECORDS FROM OTHER SOURCES TO THE FRAME

Another challenge was presented by the files from the provincial registers of nonprofit organizations. A feasibility study had shown that the inclusion of the organizations in these registers was essential for the coverage of the study.

The provincial files did not have a unique identifier such as the BN therefore the merging process to identify records from provincial files that were already on the frame from the CCRA files relied completely on the name, address and postal code of these organizations. The procedures to standardize the names and addresses, and the use of PCODE, were applied to the set of provincial records. Without the availability of the BN, a matching hierarchy allowed records from the provincial registers to match to CCRA records already on the frame according to the following order:

1. Concatenated name, parsed address, postal code
2. Concatenated name, original (unparsed) address, postal code
3. Concatenated name, postal code

With each step of the hierarchy, the certainty of a match dropped and more manual review of the matching results was necessary. Overall, the process was intended to match conservatively to avoid the loss of coverage. It was less important for duplicates to remain unmatched at the time of

the creation of the frame because it was possible for duplicates to be identified during the collection process.

For those records from the provincial files that could not be matched to the CCRA files, there were still no BN, nor an indicator of main activity. Some CCRA records were also lacking some stratification information. For these reasons, a first phase sample of records was selected to collect basic stratification information. Most records subject to this preliminary phase came from provincial files. Some respondents to the first phase were able to provide a BN number, which could be used to verify whether they already appeared elsewhere on the sampling frame. The second phase of the survey included a sample of the respondents of the first phase (after removal of duplicates that were found using the BN that had been collected) as well as the other records for which stratification information had been obtained from the frame. By using a relatively inexpensive first phase to obtain stratification information and the BN (if available), this approach led to a higher response rate and a reduced rate of duplicates associated with the second collection phase.

4. RESULTS

In the merging of the CCRA and the BR files, there was one-to-one correspondence for approximately 50 000 records because the BNs appeared once on each of CCRA and BR files. The remaining 32 000 records on the CCRA file were in fact 10 000 individual BNs, that could also be found on the BR file, but appeared more than once on either or both files. Using the Business Number and the parsed addresses to merge the multiple locations, the final result of the merge for the CCRA and BR files was a 99% match rate.

The reliability of the files of provincial registered nonprofit organizations, coupled with the fact that they had no unique identifier in common with the CCRA files, necessitated the use of hierarchical matching using the name, address and postal code fields. The use of concatenated name, parsed address and postal code of the organizations resulted in a 27% match rate. A second pass using the concatenated name, the original address and postal code of the organizations led to less than 1% match rate. Using only the concatenated name and the postal code an additional 9% match rate was achieved. However, in that last approach, some false matches were created due to the fact that in rural sectors, two different but close enough locations with the same organization's name could have the same postal code.

Another false match situation came from the incompleteness of the addresses in some files. If only the municipality - and not the street address - was provided in a certain record's address, the parsing function of PCODE would return the same string as for example "CONOTTAWA", a result that could not be detected through automated process. There were also failures to match if the names or addresses were misspelled or incomplete, or if they could be written in different ways. As an example, an organization's name like AAA could be spelled Triple A or 3A. Unfortunately, there were no automated methods to detect those differences.

5. CONCLUSIONS

The merging of the different source files for this survey was necessary to create as clean a sampling frame as possible which involved removing the duplicates and obtaining necessary information to draw a stratified sample from the frame. The quality of the source files and the availability of a common and unique identifier, even if not at the sampling level, led to variable match rates and to some false matches and failures to match.

After the merging process between CCRA files with the BR and the provincial files, the portion of the records left unmatched that did not have stratifying information was sent to a preliminary collection phase in which the BN (if available) and stratification information were collected.

In conclusion, because of the lack of common and unique identifier at the sampling unit level in the survey, some other methods for merging two records on a different set of files had to be used. These methods were useful to eliminate duplicates and to obtain additional information but they sometimes led to false matches and failures to match. Other existing merging methods could have been used, like statistical matching, but could not be considered in this survey due to time constraints.