Development of a Record Linkage System for Studying Education in Canada

Mark Krzeminski
Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A 0T6

Abstract

The Education Longitudinal Linkage Platform (ELLP) was developed at Statistics Canada to improve the ability of analysts to study education in Canada. The system allows the linkage of annual post-secondary and apprenticeship registries longitudinally and cross-sectionally and allows the addition of multiple years of fiscal data from tax records as well as possibly other administrative data sources in the future. The core of the platform is a set of keys that allows linkage of the component databases as well as the identification of individual students and apprentices across multiple years. New registries become available every year and these are added to the ELLP through a series of processing steps including linkage to tax records and identification and grouping of records referring to the same person. Analytical files produced from the ELLP enable Statistics Canada analysts to study student pathways through the education system as well as labour market outcomes. Currently, the ELLP includes education-apprenticeship data for the years 2005 to 2013 as well as fiscal data for the years 2004 to 2014. Throughout the entire process, strict attention is paid to confidentiality of the data.

Key Words: education, record linkage, fiscal data, administrative data

1. Introduction

The Education Longitudinal Linkage Platform (ELLP) was developed at Statistics Canada to increase the analytical potential of existing education databases. It achieves this by enabling the construction of analytical files through cross-sectional and longitudinal linkage of national annual post-secondary and apprenticeship registries as well as addition of fiscal data through linkage to administrative tax records. Such linkages are enabled by the creation of a person-level unique number that is consistent across all component education databases as well as the addition of Social Insurance Number (SIN), which allows linkage to tax records. To comply with Statistics Canada's strict confidentiality policies, analysts using the files do not have access to any information that could identify individuals.

This paper is organized as follows. The first section provides some background and motivation for the development of the ELLP and a description of the component databases. Section 2 explains how the platform was constructed and describes its present structure. In Section 3, the construction of analytical files is demonstrated using a simple and brief example. Finally, Section 4 provides anticipated directions for future development.

1.1 Motivation for the ELLP

At Statistics Canada, educational and social analysts are particularly interested in two aspects of education: student pathways and graduate outcomes. The first involves studying an individual's career through the education system including field of study, credentials/degrees obtained, length of time to complete a degree, whether the student graduated, etc. The second involves studying the labour market activity of an individual after he or she has graduated to determine length of time to find employment after graduation, annual income, income trend, whether or not an individual received social assistance, etc.

Many studies of post-secondary education based on data from sample surveys of individuals have been conducted at Statistics Canada in the past, such as the National Graduates Survey and the National Apprenticeship Survey, which are repeated at irregular intervals. However, these were cross-sectional studies. The most recent longitudinal sample survey of individuals on education was the 2002/2005 edition of the National Graduates Survey / Follow-up Survey of Graduates (NGS/FOG). This project targeted the graduating class of 2000: a sample of graduates was selected for the NGS in 2002 and respondents were interviewed a second time in 2005 for the FOG. The 2007 and 2013 editions of the NGS did not involve follow-up surveys. Thus, there was no recent sample survey on which to base a longitudinal study. Conducting a longitudinal sample survey would have disadvantages such as difficulty reaching individuals – especially for the highly mobile population of recent graduates, the possibility of low response rates, a high attrition rate, and high respondent burden. Thus, attention turned to existing administrative data sources.

1.2 Component Databases

There are three databases that are part of the ELLP: the Postsecondary Student Information System (PSIS), the Registered Apprenticeship Information System (RAIS), and the T1 Family File (T1FF), which are described below.

The Postsecondary Student Information System is an annual registry of enrolments in all Canadian public postsecondary institutions during the academic year. It collects information pertaining to the individual student, such as name, date of birth, gender, address, and academic status, as well information pertaining to the program in which the student is enrolled, such as field of study, program level, and duration. An annual PSIS file typically contains around 3 million records. The data are provided to Statistics Canada either by the institutions directly or by provincial ministries of education. There is often considerable variation in the completeness of the information provided – certain institutions and provinces do not even provide a name and date of birth – which must be taken into account when linking to administrative records to obtain a Social Insurance Number.

The Registered Apprenticeship Information System is a national annual registry of trade apprentices taking in-class or on-the-job training. As with PSIS, records contain information pertaining to the student and the program. An annual RAIS file typically contains just under half a million records.

There is a considerable overlap of individuals on both files since PSIS includes trade-vocational training programs as well as any in-class portion of an apprenticeship program, and since the distinction between a trade and an apprenticeship differs somewhat from province to province.

The T1 Family File is a database developed and maintained at Statistics Canada, which is derived from income tax declarations and other administrative files provided by the Canada Revenue Agency. For a given tax year, the T1FF includes information on income from various sources, field of employment, tuition and education deductions, government transfers, as well as some demographic and address data. The linkage to the T1FF is done by matching on Social Insurance Number (SIN).

At the time of this writing, the ELLP includes the PSIS files for 2008-2013 (also 2005-2007 for the Atlantic provinces), RAIS files for 2008-2013, and tax files for 2004-2014. Other years will be added as they become available.

2. Construction and Structure of the ELLP

In this section, we describe how the ELLP was put together and how new files are processed and added to the platform.

Essentially, the ELLP consists of three elements:

1. A set of keys enabling cross-sectional and longitudinal linkage of component databases. The keys are MASTERID, which is a unique person-level number derived by processing the files and is consistent across all education databases, RECID, which is essentially the line number on a particular education database, and SIN (Social Insurance Number), which is a unique person-level identifier on tax databases.
2. A procedure to add new databases, both education and tax, to the platform as they become available. The procedure allows the correction of existing information on the platform based on information in an incoming file.
3. A procedure to produce anonymized analytical files according to specifications provided by analysts. Such files do not contain any personal identifiers such as name, date of birth, etc., or any Social Insurance Numbers.

Each file was processed and added separately according to the procedure described below.

2.1 Pre-processing of a new file

Upon receiving a new PSIS or RAIS file, some pre-processing is required to prepare the file for addition to the ELLP. Many records on the PSIS and RAIS files already have a SIN but not all of them. In order to obtain SINs for as many records as possible, after a line number (RECID) is assigned, the file is linked to an administrative tax database called the Linkage Control File (LCF), which contains Social Insurance Number and personal identifiers as they appear on historic tax returns since 1981. The linkage to the LCF is done by a series of hierarchical deterministic linkages on personal identifiers. It may also happen that some individuals, such as non-permanent residents who became permanent residents, or individuals who have been victims of identity theft, have had multiple SINs during their lifetime, which all need to be taken into account when linking their records in education files to their tax information. To this end, a cross-reference file is used to obtain all SINs belonging to an individual.

Average yearly linkage rates for this process are around 96% for RAIS, while about 63% for PSIS, due to many records on the latter missing key identifiers such as name and date of birth. If only records having both name and date of birth are considered, the average linkage rate for PSIS increases to 84%, which is a substantial improvement. The SINs that were already available in either PSIS or RAIS are used to estimate linkage error rates. Average yearly estimated false positive and negative rates are 0.3% and 8.4% respectively for PSIS, and 0.2% and 3.3% respectively for RAIS.

The second step is to identify and group sets of records on the file pertaining to the same individuals. Since these files are lists of all registrations, not individuals, a particular individual may appear more than once on a particular file. This may happen because an individual was enrolled in multiple programs simultaneously or due to an administrative particularity or error at an institution. This identification is done by a linkage of the file to itself ("internal" linkage) by matching on institution and student number, as well as personal identifiers in order to find individuals enrolled simultaneously at different institutions. Groups of records pertaining to the same individual are assigned the same MASTERID.

The third step is to split the linked file into two:

1. "identifier" file: contains RECID and personal identifiers such as name, date of birth, etc. as well as province of institution, institution code, and student ID;
2. "education" file: contains RECID and "education" variables pertaining to the student's academic status and the program in which he or she is enrolled;

This split is done to comply with Statistics Canada's strict record linkage policies to ensure confidentiality of personal data. Access to the linked files is restricted to the few persons responsible for maintenance-development and production of analytical files.

2.2 Initial creation of the ELLP

After addition of SINs and MASTERID and separation of identifiers from education variables, the file is ready for addition to the ELLP. The construction of the ELLP began with the 2010 PSIS and RAIS files, which were the most recent vintages available at the time.

First, a "Keys" file was created from the pre-processed RAIS 2010 containing only the RECID, MASTERID, and corresponding SINs[1]. Then an "Education Master File" (EMF) containing corresponding MASTERID and personal identifiers was also created. The "Keys" file and EMF formed the seed of the ELLP, to which other files would be added one at a time.

Second, the pre-processed PSIS 2010 file was linked to the "Education Master File" by a series of hierarchical deterministic linkages on SIN, institution, student ID, and personal identifiers. In this way, individuals common to both files were identified and MASTERIDs reassigned to make them consistent across both files. A corresponding Keys file was created for PSIS 2010 and personal identifiers were added to the EMF.

[1] An individual may change SINs under special circumstances, e.g., following naturalization, or in response to identity theft.

2.3 Addition of subsequent files to the ELLP

The procedure for adding is essentially the same as the original creation of the ELLP. All subsequent files, i.e. PSIS 2005 to 2013 and RAIS 2008 to 2013, were pre-processed and added to the ELLP one at a time as described above, and corresponding Keys files were created and MASTERIDs were reassigned on the incoming file to preserve consistency. Records for which we were unable to find a SIN were still added to the ELLP, as these records can still be used to study student pathways.

2.4 Updates and corrections to the ELLP

Addition of a new file to the ELLP brings new information that can be confronted with existing information. Such a confrontation can reveal errors either on the ELLP or on the incoming file. Such errors are usually due to partially missing information that is completed by the incoming file. There are two types of linkage errors:

1. False matches: A set of records may have been grouped as pertaining to a single individual while in fact they pertain to two different individuals. For example, two records already on the ELLP may have very similar names, one of the records is missing postal code or phone number, and the same SIN was wrongly assigned to one of the records by the educational institution. The incoming file may have complete postal code and phone number for the two records, strongly suggesting that the records in fact belong to two different individuals.

2. Missed matches: A group of records that in fact pertain to one individual may not have been identified as such and have thus been assigned two different MASTERIDs. This may happen if the data on one or both of the original records are missing, erroneous, or have changed. For example, a woman getting married may change her surname as well as postal code and phone number at the same time and so records from before and after her marriage may look quite different.

When errors are identified, they are corrected, which involves updating the affected Keys files by changing MASTERIDs or by deleting SINs. Usually, these corrections involve a small proportion of a file (less than 0.1%), but this can still represent thousands of records. Automated procedures have been developed that address the majority of these cases, but some manual verifications and possible corrections are still required.

3. Example: Extracting a Longitudinal Analytical File

We now present a simple example to illustrate the procedure for producing a set of analytical files that are linked longitudinally. Suppose an analyst wishes to study individuals enrolled in post-secondary education in Canada in all of the years 2010, 2011, and 2012 and to have their tax data for 2012. To produce the corresponding set of files, we start with the Keys files corresponding to PSIS 2010, 2011, and 2012 and identify the MASTERIDs common to all three files. Table 1 shows simplified versions of what keys files for the years of interest look like.

Table 1: Example of Keys files for three years of PSIS

Keys for PSIS 2010		
RecID	MID	SIN
L11	M01	101
L12	M02	102
L13	M02	102
L14	M03	103
L15	M04	104
L16	M05	105

Keys for PSIS 2011		
RecID	MID	SIN
L21	M06	201
L22	M07	202
L23	M02	102
L24	M08	203
L25	M04	104
L26	M05	105

Keys for PSIS 2012		
RecID	MID	SIN
L31	M09	301
L32	M02	102
L33	M10	302
L34	M11	303
L35	M04	104
L36	M12	304

We see that the only MASTERIDs (or MID above) common to all three files are M02 and M04. These records are indicated in grey. Thus these are the individuals of interest. The corresponding RecIDs are then used to fetch program and socio-demographic data from the corresponding PSIS file and the corresponding SINs are used to fetch fiscal data from the 2012 tax file. Furthermore, if requested, some indicators can be derived from the linkage process and added to the files. The person-level key, MasterID, is retained on the resulting analytical files, but is masked as "RequestID". The resulting files resemble those shown in Table 2.

Table 2: Example of analytical files delivered to analysts

PSIS 2010			
RequestID	Education data	Tax data 2012	Indicators
R01	…	…	…
R01	…	…	…
R02	…	…	…

PSIS 2011			
RequestID	Education data	Tax data 2012	Indicators
R01	…	…	…
R02	…	…	…

PSIS 2012			
RequestID	Education data	Tax data 2012	Indicators
R01	…	…	…
R02	…	…	…

We note that these files that the analysts work with do not contain any personal identifiers such as names, etc. and they do not contain any Social Insurance Numbers.

4. Future Development

Regarding the linkage to tax records to obtain Social Insurance Numbers, it is planned to switch to a probabilistic linkage methodology, for which Statistics Canada has sophisticated software, called G-Link. Also, linkage of the education databases to other files is projected, such as employment insurance records, the national immigration database, and student loans files. Furthermore, it is planned to eventually integrate the ELLP with Statistics Canada's *Social Data Linkage Environment*, which currently integrates labour, income, justice, health, and vital statistics.

Acknowledgements

I wish to first thank the members of my methodology working unit: Kim Bornais, Dany Faucher, Carole Morin, and Martin Pantel, who undertook the majority of the work in building the record linkage system described in this paper. I especially wish to thank Martin Pantel, on whose work this paper is substantially based. I also wish to thank our internal clients at Statistics Canada, the team of analysts and subject matter experts: Sylvie Gauthier, Karen Griffith, Christine Hinchley, Aimé Ntwari, Anna Rigakis, and others.

References

Caron, P., Ellison, J., Faucher, D. (2013). "Linkage of education files", unpublished document, Ottawa, Canada: Statistics Canada.

Pantel, M. (2016). Using administrative data to study education in Canada. *Proceedings of the 2016 Statistics Canada Symposium*.