Work report :

Summary of the stages of the setting up of the GGP contextual database for Canada

Preliminary version

Produced by :

Marc-Antoine Busque

Patrick Charbonneau

Yann Décarie

Guillaume Marois

Département de démographie

Université de Montréal

Thursday August 17, 2006

Introduction

If two out of the four members of our group have been in contact, beforehand, with the *Generations & Gender Programme* (GGP), through some exploratory works they have made, it nevertheless appears that our actual mandate really started up at the beginning of May. This mandate consisted in setting up, for Canada, the GGP's Contextual Database, by finding the data required for the variables it includes.

Before going further, we would like to say a few words about the way the GGP was conceived. This programme is made of two major components, totally independent from each other at the data gathering level, but that could be interactive at the statistical analysis level: the *Generations & Gender Survey* (GGS) and the *Contextual Database* (CDB). The GGS consists of a panel survey of three waves (three years apart) in which 10 000 individuals aged from 18 to 80 are followed. The CDB, on the other hand, relates to more than 200 variables, of national and/or regional level, sometimes qualitative but more frequently quantitative (time series from 1970 up to present in most cases), related to a wide range of topics: health, economy, employment, culture, education, demography, pensions, etc. It's toward the making of that second component of the GGP that was oriented our mandate.

Support received

Before we concretely began to seek for data, we have had recourse to some theoretical and technical support. As regards the theoretical, the conceptual aspect, we could impregnate ourselves of it through the reading of many documents made accessible via, *inter alia*, the Population Activity Unit website which contains papers from the Informal Working Group, the Consortium Board and the GGS Expert Working Group. Thus, we could understand the conceptual aims that stand behind the choice of variables that had been made and to seize the bonds which linked the CDB with the GGS.

As regards the support for the more practical, the more technical aspects of our task, we must recognize that the collaboration of Mr. Martin Spielauer and his team from the Max Planck Institute for Demographic Research (MPIDR) largely facilitated our work. From the beginning, we were entitled to a whole series of Internet links towards international data bases containing information related to our variables. In addition, Mr. Spielauer also provided us all the "templates" of Bulgaria. In this manner, we could see how they had laid out their variables in the Excel spreadsheets, the sources they had used, etc. Finally, we also received a list of all the variables with their definitions. In short, at this point, the only thing left to do was to plunge…

Seeking for variables

Knowing that one of the goals of the CDB is to compare the participating countries, we believed that it was preferable to begin our research of variables by the international databases. In fact, if all the countries are doing the same, it will be easier to make adequate comparisons. To help us in our step, a list with a short description of this type of databases was put at our disposal, thus facilitating the beginning of our research. Quickly, we realized that it was not so easy to find complete annual time series (1970-2006). Moreover, the regional aspect, the provinces with regard to Canada, is seldom taken into account by international databases. Also, there are several databases which are devoted only to European data with certain variables that we do not find at the international level for the other countries. To mitigate these small problems, we directed our research towards national databases, for example, towards the data of Statistics Canada. This enabled us to find several variables which did not seem to be taken into account for Canada by the international databases. Moreover, it is easier to find data of regional level on this type of databases. However, the difficulty to find complete time series remains present. Let us note that the recourse to the national data does not constitute, however, the ideal option since the statistical standards can diverge from a country to another.

Our searches were mainly carried via Internet. Although this support represents a very great source of easily accessible information, we were not able to make it completely without the old paper sheets. That proved to be necessary in particular because we had to find several data

from a quite distant past, and it seems that the government finds expensive to digitize all information available through time…

The website "Sherlock", shared infrastructure for the management and distribution of survey data, were largely useful for some specific variables. However, as they are very small samples and to preserve confidentiality, the categories of some variables were not exactly such as we wished them. Moreover, technical difficulties were frequently encountered with the system and several of them appeared insoluble after consultation of the person in charge for Sherlock at UofM.

"Textual" variables

An original characteristic of the CDB consists of inclusion, by its creators, of several qualitative variables, of "textual" variables. It is undeniable that this aspect of the CDB will be of a great utility for the researchers, allowing them to better catch the context in which fits the phenomena they study in order to set up assumptions and to draw conclusions fastening themselves more narrowly with "reality". It appears, nevertheless, that the search for these variables gave us some difficulties.

One of the first problems we had to face concerns the sources of information. It proved to be difficult to find, for several variables, reliable sources providing all the information required in the text holding place of variable. Thus, the necessity to look at various sources for only one variable let us, in many cases, in the doubt about knowing if we had on hand all existing information. It was then necessary for us to seek more in-depth, to rake the contents of as many sites treating of the various variables as possible, in order to compare our various sources and to make sure that no information had been left behind.

The difficulty with these qualitative variables rises mainly from the fact that a large majority of these "textual" variables consist of a description of the evolution since 1970 of various "systems" (health, education, pension,…) and that, if it is relatively easy to describe the

current operation of those, the information on the modifications which they underwent in the course of time remains difficult to reach. It is hard to know which measurements, which laws preceded those in force at present time. The governmental sites, for the majority, do not draw up the history of these changes. Thus, we had, as we mentioned, to diversify our sources of information (emails and phone calls to the organizations concerned, research on non-governmental sites, etc).

Lastly, always regarding our search for these variables, but also for some variables of quantitative nature, we met some obstacles related to the political structure of the Canadian confederation. Indeed, some services being of federal competence, others of provincial jurisdiction and others being divided between two levels of governments, we have had to modify our way of structuring information according to the type of jurisdiction which prevailed. For example, when we had to describe the educational system, which is of provincial competence, we could not resort to only one long description for the whole country since notable differences exist from one province to another. On the other hand, we could not either describe the system of each province, which would have been too much tiresome. Thus, we had to emphasize the elements common to each of these systems, while mentioning what was specific to each one of them. Once again, obviously, the whole problem of "history" comes back. Where can we find the information related to the evolution of the educational system of each province since 1970? We are still working on it…

Resources

Concerning the research of data, to obtain other information or for being better oriented in our research, we contacted various departments, organizations or key people who could have information on some of our variables. Among those, there are, among others things, the various branches of Statistics Canada, the Department of Social Development, the Council of Ministers of Education of Canada or the International Labour Organization (ILO).

We proceeded in two ways: sometimes by phone, but most of the time by email. To reach a better coordination, we created an email address for the team with the name of "Équipe Vieillissement". To write our emails, we followed a model-letter explaining our project, in which we changed the information requested according to the correspondent.

If we know beforehand key people working in some of the organizations concerned and whose interests of work agree with some of our variables, we initially tried to contact them. For example, we directly sent email to Laurent Martel, former student of the "Équipe Vieillissement" in the Department of Demography, asking for some information on the elderly living in institution. Otherwise, our emails were generally sent for the general email address of the organizations. The requests by phone call were all directed to the services of general information, which thereafter were transferred if there was a need.

Although some queries were successful, the majority of them did were not answered or had negative or unsatisfactory answers. On about thirty requests which were sent, a dozen could be fruitful, either by providing us the totality of our request, by answering them partially, or by referring us to other contacts. For the other requests, the answers were either negative (no data, not in their competences, etc), or at cost or there was simply no answer.

We can thus conclude that the cooperation with some organizations was not very successful. Several data considered as missing exist, but we could not have access to them directly at no cost.

When we had more technical questions on the database (e.g. problems of definition), we were told to contact Mrs Dora Kostova, from *Center for Population Studies of the Bulgarian Academy of Scienc*e. She was a collaborator of the project and involved with the Bulgarian database. Unfortunately, she was on holydays until June 6 and took a certain time before coming back to us. We had to wait until June 19 to have a reply to our May 30 message. In spite of that, the answers were satisfying and our interrogations were solved.

Resolving our problems

During our searches, we have more than once encountered some problems in order to follow the rules set by the templates. These difficulties mainly occurred when we had to find variables that were not accessible through well known international databases as those of the UN, OECD and UNESCO. In fact, the lack of standardized data forced us to consider data coming from national organisations, such as Statistics Canada, which was often consulted, even if these data were not always normalized with an international perspective. Moreover, Statistics Canada (and other non-international organisations) data are sometimes presented in an other format than what we would have liked. Thus, we had to take some decisions in order to find accommodations that would allow us to complete a variable of the contextual database even if the splitting of the categories and/or even some definitions did not exactly match between the data collected and what was required in the templates. For example, variables 1204a and 1204b were requesting the highest educational attainment in accordance with standard schooling levels (*ISCED, International Standard Classification of Education*). However, the only data gathered were coming from Statistics Canada and were not classified in function of the ISCED system, but rather in accordance of the highest diploma obtained (High school graduation certificate, Trades certificate or diploma, Bachelor's degree, etc.). We then had to restructure the classes to make them accurately comparable with those that were required. In other cases, we did not have other choices than to keep some categories as they were even if they were not exactly respecting the instructions given in the templates. We experienced this particular problem especially when the data must be provided by age groups: more than once, the age groups for which the data were available were not subdivided as we would had liked (for example, we sometimes only had access to data compiled in 10 years age groups whereas we would have need 5 years age groups). When some changes required to be commented in order to be better understood, footnotes were incorporated in the concerned tables. Additional comments were also added when some definitions of concepts needed further explanations, ether because they were more complex or because they were not considered as standardized in an international perspective. Furthermore, some tables were subdivided in two separate parts so that they could adequately respond to a more various range of eventual requests. For some variables, as labour-market participation, two distinct tables were realized; one showing the labour-market force in absolute numbers and the

other one presenting labour-market participation rates. This was done even if only the data concerning absolute numbers of labour-market force was required. In other cases where some variables were more complex in the case of Canada than in the rest of the world, tables sometimes also required to be subdivided in two parts. Variables concerning marital status are good examples to illustrate this situation, as we had to produce two sub-tables, one concerning the legal marital status and the other concerning marital status where people living in common-law were considered as married. Once more, this was done with the objective to answer to the most various possible ranges of eventual requests.

Overall, the majority of the data was collected without considerable difficulties; and the definitions, the general presentation of the data and their categories were most of the time compatible with the standards requested in the templates. When it was not the case, we had to agree to establish some compromises, which have always respected the real signification that had be described by the variable in question. Everywhere where we had to add some modifications, we have taken the precaution to describe at our best the new changes, while being brief, clear and accurate.

Conclusion

Thus, we practically found all variables which are accessible at no cost by Internet or in the statistical directories to which we have access. The majority of the missing variables or the incomplete parts remaining are non-existent, at cost or inaccessible for us. The database is thus not complete and it would be a utopia to think that it could be complete. The fact that there are no internationally standardized data, that it is difficult to find the chronology of some variables, that not all the templates are adapted to Canadian reality and that some variables are simply non-existent for Canada are some of the factors explaining this situation. However, the database in its current status is full of relevant information being useful to pursue the project.