

Administrative and Survey Files for Research Purposes

Line Beauchesne

Information Integration and Technologies, Institut de la statistique du Québec
line.beauchesne@stat.gouv.qc.ca

Abstracts

Existing data banks in Québec are a gold mine for those who are interested in developing programs liable to improve the health and welfare of the population. It is essentially because they are difficult to access that their use is limited and that researchers, analysts or public decision-makers are led to undertake costly and often lengthy studies involving the collection of often already available data.

That is why the Environnement pour la santé et le bien-être (EPSEBE – [Environment for the Promotion of Health and Welfare](#)) project, a centre of expertise and an infrastructure for the processing of information, aims to offer new methods. It notably provides a portal of specialized services, including, among other things, a dictionary with standardized definitions of metadata including cautions regarding their use; the possibility of sharing research results and indicators; access to a team of experts ready to advise and assist requestors in the preparation of their research projects; the pairing, with or without a unique identifier, of administrative and survey files required by the research project; and the remote use of administrative and survey files, using standard tools and under ethical and highly secure conditions, for research or analysis purposes.

The purpose of the presentation is to present the project as well as the various services, tools and quality principles offered by the service platform.

Keywords: metadata, linkage file, remote access

1. Introduction

The Institut de la statistique du Québec (ISQ), the official statistical agency of the province of Québec, has been offering since July 2007 a remote access service in order to provide access to research files resulting from linkages.

After a portrait of the background of this project, this presentation will give a general description of the project: the submission of the request by the researcher, the processing of the request, the remote access to the research file, and the environment and the development of the service platform. Then some of its added values will be presented. Lastly, the current status of the project will be outlined.

2. Background

In recent years, the ISQ has been faced with an increasing demand for statistical information and a growing diversity of users. Users, particularly researchers, are increasingly seeking access to microdata files that they can process themselves.

To enable us to respond adequately to such needs, steps have been taken in various fields, in accordance with our strategic orientation of supporting research in Québec. Following the setting up in the year 2000 of our research data centre where researchers go in order to access ISQ data, we innovated by setting up a service platform which provides highly secure remote access to research data. Initiated by researchers for researchers from university circles or public agencies, this new service is intended for all disciplines.

In 2004, the ISQ started a partnership project with a network of researchers and other government agencies. Through this partnership, we undertook to set up a new service platform.

Existing data banks in Québec are a gold mine for those who are interested in developing programs liable to improve the health and welfare of the population. It is essentially because they are difficult to access that their use is limited and that researchers, analysts or public decision-makers are led to undertake costly and often lengthy studies involving the collection of often already available data.

Confidential information from administrative files can be transmitted to researchers in keeping with Québec's privacy protection legislation. However, the disclosure of confidential information is subject to rigorous oversight: Québec's Access to Information Commission must first give its authorization, and stipulate the terms of use and destruction of the information.

Access by researchers to confidential information taken from administrative files is a common and relatively simple process when the information is requested from a single holder agency. However, the process becomes much more complicated and long delays can be experienced if the researcher needs access to files belonging to different holders in order to link them. Moreover, the linkage of files originating from different holders or domains, such as health and education, becomes a quite complex statistical operation. This is because, in Québec, there is no unique personal identification number. Consequently, these linkages should be done with probabilistic methods on the basis of names, addresses or other available variables.

Because of our legal oversight, our normative framework for the protection of personal and confidential information, and our statistical expertise in databank processing, the ISQ has been chosen to host the service platform.

3. General description

It is important to point out, from the start, that the service platform is not a data warehouse. Rather, it makes it possible to manage a set of requests through which only authorized researchers have remote access, during the time allowed for each research project, to research files resulting from the linkage of other files.

The service is accessible by means of an Internet portal, through which researchers can submit their requests, follow up on them, and make remote use of the research files. The portal of specialized services includes, among other things, a dictionary with standardized definitions of metadata, including:

- cautions regarding their use,
- the possibility of sharing research results and indicators,
- and access to a team of experts ready to advise and assist requestors in the preparation of their research project.

This dictionary is very useful for researchers in both the exploratory and analysis phases of their projects.

The way this service platform works follows.

3.1 Submission of researcher's request

First of all, the researcher must submit a request for access. To help the researcher formulate his request, there is a search engine on the Internet portal that can consult the dictionary in order to obtain information on a specific data source or a holder of information. It is also possible to search by keyword.

On the basis of his protocol, the researcher uses the request form on the Internet portal to submit his demand then automatically directed to the ISQ.

3.2 Processing of researcher's request

Then the researcher's request is processed and analyzed.

The first processing stage consists in creating the research file, that is, the file that will result from the linkage of the administrative or survey files mentioned in the researcher's requests. The exchange of data between the ISQ and the holders is carried out in such a way that, during a given exchange, neither the identification data nor the content data are available at the same time.

Record linkage is a process of bringing together two or more separately recorded pieces of information belonging to the same entity. Probabilistic linkage does not require complete agreement on the matching variables. Instead the quality of the match between pairs of records is rated using sophisticated algorithms to evaluate the likelihood of a correct match between two records. Generalized Record Linkage software (GRLS), developed at Statistics Canada, is used by the EPSEBE's statisticians to execute the linkage. In addition, for estimates produced from survey data to be representative of the target population, and not just the sample itself, survey weights are used. With record linkage between administrative files and

survey files, additional survey weights might be required because not all survey respondents can be linked to their administrative data.

The second processing stage consists in controlling the risk of disclosing confidential information in the file by applying appropriate masking techniques.

3.3 Remote access to research file

Finally, the researcher accesses the research file. High-level security measures for authentication of a researcher authorized to access the service platform are carried out with an identifier and a personal password, reinforced by the use of a token given to each authorized researcher.

As it is not possible for the researcher to download or print the data or results, when he wishes to recover the results of his analyses, he must submit the request through the Internet portal. Our specialists examine the results in order to control the risk of disclosure in accordance with the ISQ's guidelines on the subject. When they are in compliance, we route them to the researchers.

Lastly, the user's work session is recorded and can be studied at any time, if there is a need to confirm the integrity of the researcher's operations in order to ensure that confidentiality and security are abided by.

3.4 Environment and development

The environment and development of the platform are based on Oracle technologies and free software. The software used is listed below:

- Products:
 - Oracle Fusion Middleware 10G (AS and BPEL, Collaboration Suite, DB)
 - Suze Linux
- APIs:
 - Oracle ADF, Oracle BPEL, Jasper Reports, JAAS, XML, PL/SQL
- Development tools:
 - Oracle Jdeveloper, Tortoise, DMS Loading Tool

4. Added values

The service platform presents several added values. Most of all, it is an innovation in Québec. It meets an obvious need to place at the disposal of researchers files resulting from the linkage of administrative or survey files. Now that record linkage is a well-established technique in population studies much can be learned about social, economic and health determinants or outcomes, and so on.

The new service facilitates research, puts new means at the disposal of researchers, and is expected to accelerate the current access process. It also includes an

economy of scale and the possibility to save work when it comes to retrieving and processing information. It makes possible the networking of knowledge and expertise, and, finally, the standardization of access to information requests.

The location of the service platform at the ISQ provides guarantees that the confidentiality and security of the data will be ensured, given our legal oversight and normative framework in that regard.

So, the researchers involved will have access to their research files more quickly. More importantly, the use of the platform, gives Québec's Access to Information Commission and agencies holding administrative files additional guarantees of privacy protection, since no nominative information is accessed by the researchers.

5. Current Status of the Project

The service platform has been in operation since July 2007. While some researchers's requests are currently being processed and analysed, we are working on different axes to consolidate the project.

We are intensifying the promotion of the project to enlarge the field of users. We are also in the process of negotiating agreements with our different partners, that is, the major agencies holding administrative files and Québec's Access to Information Commission, in order to agree on a processing time limit. And we are preparing a business plan that will include, among other things, our pricing policy, and our action plan for the next five years.