Distr.
GENERAL

Working Paper No.12
11 April 2007

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2007)
(Geneva, 8-10 May 2007)

Topic (ii): Statistical information systems architecture

REMOTE ACCESS TO RESEARCH FILES RESULTING FROM LINKAGES

Supporting Paper

Prepared by Annie Giguère and Madeleine Filion, Institut de la statistique du Québec, Canada

I. SUMMARY

1. The Institut de la statistique du Québec (ISQ) is the official statistics agency of the government of Québec, one of the provinces of Canada. As such, it has taken the strategic orientation of serving the research community. Researchers regularly conduct statistical analyses and research on the basis of administrative files from Québec government agencies, or survey files. It is relatively easy for researchers to access confidential information from a single holder agency. However, the situation becomes a great deal more complicated when they seek access to files belonging to several holders in order to link them, particularly since the files often do not have a unique identification number, which could facilitate linkage.

2. To promote research, in keeping with privacy requirements, the ISQ initiated a project in 2004, in partnership with a network of researchers and government agencies, aimed at setting up a platform for research services that would provide an individual, secure, remote-access working environment for each research project and maintain the confidentiality of personal information. The goal of the project is to provide the scientific community with a tool and support for the use of research files resulting from the linkage of various administrative or survey files held by public agencies. The ISQ will host the service platform.

3. Initiated by researchers for researchers, this new service is for researchers in all disciplines, whether from university circles or public agencies. After several European experiences, the platform for research services is a major innovation in Québec and meets a definite need. It also provides researchers with:

- an Internet portal offering specialized services providing remote access under ethical and highly secure conditions to research data of interest;
- a team of experts to advise and accompany researchers in preparing their research request and processing their file.

4. Design, building and testing are to be completed by the spring of 2007 and the new service platform will be in operation as of the summer of 2007.

II. INTRODUCTION AND BACKGROUND

5. In recent years, the Institut de la statistique du Québec (ISQ), the Québec government's official statistics agency, has been faced with an increase in demand for statistical information and the growing diversity of users. Users, particularly researchers, increasingly seek access to microdata files that they can process themselves. To enable the ISQ to respond adequately to such needs, steps have been taken in various fields, in accordance with its strategic orientation of supporting research in Québec. Not only has the ISQ acquired a normative framework for maximizing access to its statistical information by third parties while respecting privacy, but it has also undertaken a partnership project with networks of researchers and other government agencies in order to maximize use of the potential of existing data sources in Québec's public administration.

6. In 2004, through this partnership, the ISQ undertook to set up a service platform that, when fully operational, will give researchers secure remote access to research files resulting from the linkage of various administrative or survey files already available in public agencies.

7. In Québec, many administrative files are accessible, under various conditions, for research purposes. Among the most in demand are health insurance files, hospitalization files, birth and death registers, and files on educational progress, to name but a few. Confidential information from administrative files can be transmitted to researchers in keeping with Québec's privacy protection legislation. However, the disclosure of confidential information, particularly personal information, is subject to rigorous oversight: the Commission d'accès à l'information du Québec (Québec's access to information commission) must first give its authorization, and stipulate various terms of use and destruction of the information.

8. 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. 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.

9. This situation can be explained as follows:

- Very often, the files to be linked do not have a unique identification number. Therefore, the linkages must be carried out on the basis of identifying variables (name, address, date of birth, etc.), which is extremely complex and long.
- It is also difficult for researchers to obtain detailed information from file holders concerning the availability of the data and their quality.
- Researchers sometimes encounter technical difficulties, since not all the files are recorded on compatible media.
- The multiplicity of holder agencies increases the number of authorizations to be obtained and extends the time frames accordingly.
- Lastly, the public, elected officials, media, leaders, etc. are very sensitive to the linkage of files.

10. In short, the information access process resulting from file linkage is long and laborious, and not always successful in terms of the quality of the end result and the data obtained. These obstacles inhibit, and even discourage, some researchers, who turn to other areas of research, as they are obliged to act quickly in order to remain competitive in their field. Research is therefore less competitive, and those evaluating medical or social programs are deprived of relevant information. To remedy the situation, a group of researchers decided to launch an innovative project.

11. Because of its legal oversight, its normative framework for the protection of personal and confidential information, and its expertise in databank processing, the ISQ was chosen to host the service platform. A number of government agencies joined the partnership as holders of administrative files and content specialists.

12. After this backgrounder, the reader will find a general description of the project. This will be followed by an explanation of its added value. Then, the current status of the project will be outlined. Finally, the status of the work is given in order to ensure the financing and continuity of the project.

III. GENERAL DESCRIPTION OF THE PROJECT

13. It is important to point out, at the outset, that the service platform is not a data warehouse, but makes it possible to manage a set of requests pursuant to which only authorized researchers have remote access, during the time allotted to each research project, to research files resulting from the linkage of other files.

14. The service is accessible by means of an Internet portal, through which researchers can consult a data dictionary documenting the variables available, submit their requests, follow up on them and make remote use of the research files.

A. Environment and Development

15. The platform service is built on Rich Internet Applications based on SOA using Oracle tools.

16. The products used are:

- Oracle Fusion Middleware 10g r3 (10.1.3.1)
  - Oracle AS (two instances):
    - Public/researchers web application
    - Employees web application
  - Oracle BPEL
    - Batch use cases
    - Business process orchestration
- Oracle Fusion Middleware 10g 10.1.2
  - Oracle Internet Directory
    - Security
  - Oracle Metadata Repository
    - Central management of OAS
- Oracle Collaboration Suite 10g
  - Email server
- Oracle DB 10g r2
  - Application data (one instance)
  - Batch scheduling
- Suze Linux
  - Operating system (all products)

17. The APIs used are:

- Oracle ADF 10.1.3.1
  - Java/Web applications
- Oracle BPEL 10.1.3.1
  - Batch use cases
- Jasper Reports (Open Source)
  - Reports
- Freemarker (Open Source)
  - Email templates
- JAAS
  - Authentication and authorization
- XML Beans
  - Loading configuration CML files into JavaBeans components
- PL/SQL
  - Batch business logic
  - Email common service

18. The development tools used are:

- Oracle JDeveloper 10g 10.1.3.1
  - Java/Web/batch development

    - BPEL modelling
    - PL/SQL development
- SubVersion 1.3.0
    - Configuration management
- Tortoise SVN 1.3.5
    - Management of Java source files in Windows Explorer
- DMR loading tool
    - Test data

B. How the Service Platform Works

(a) The Researcher Submits a Request

19. To help the researcher formulate the request for access, on the Internet portal is a search engine 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.

20. On the basis of his or her protocol, the researcher submits the request for access to a research file using the request form on the Internet portal.

21. Once submitted, the request for access is automatically directed to the ISQ, which does not claim to judge the relevance or scientific quality of the request, or support or ensure the funding of the research protocols submitted. These questions must have been resolved beforehand with other authorities, for example, peer committees or grant agencies.

22. On the Internet portal, the researcher can follow up on the request at any time through the request module.

(b) The Researcher's Request is Analysed

23. An initial analysis of the request is conducted with regard to the availability and quality of the data requested, and the technical feasibility of proceeding with the linkages, as well as the impact on the quality of the linkages and the masking to be done. The required adjustments are made to the request, in cooperation with the researcher and the content specialists, i.e. holders of the administrative files.

24. When personal information is sought, the request is forwarded to the Commission d'accès à l'information for approval and, subsequently to the agencies holding the files involved in the request, once the Commission has given its approval.

25. Acceleration of the current access request process for research purposes is expected. The implementation of a process recognized by all parties (researchers, file holders, the ISQ, the Commission d'accès à l'information) that guarantees secure, normalized processing should be instrumental in achieving that objective.

26. Once the approvals are received, a contractual agreement is entered into between the ISQ and the researcher, for each research protocol submitted.

(c) The Researcher's Request is Processed

27. The first processing stage consists in creating the research file, i.e. the file that will result from the linkage of the administrative or survey files mentioned in the researcher's request. This is a dual stage. 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 (used for the linkages) nor the content data (required by the researcher for his or her protocol) are available at the same time.

28. Thus, initially, only the identification data used for the linkages are transmitted by the holders to the ISQ. Each record is accompanied by a distinct, anonymous sequential key created by each supplier.

29. The linkage is done on the basis of a methodology that yields the best results, particularly in terms of quality: probabilistic linkage with direct identifiers (family name, given name, address, date of the birth, etc.) or deterministic linkage on the basis of indirect identifiers.

30. The identification data used for the linkage are destroyed by the ISQ.

31. Secondly, the sequential keys for each record located on the file resulting from the linkage are returned to the respective holders. The holders then supply only the ISQ with the research data coupled with the sequential keys. All exchanges with data holders are secure.

32. The research file resulting from the linkage is then structured, and only the variables required by the protocol are entered. The second processing stage consists in controlling the risk of disclosing confidential information in the file by applying appropriate masking techniques.

(d) The Researcher Access the Research File

33. A work zone is created for each research request and becomes accessible, but strictly to people duly identified in the agreement signed between the ISQ and the researcher, and only in the areas set aside for that purpose.

34. The authentication of a researcher authorized to access the service platform is carried out with an identifier and a personal password, reinforced by the use of a token given to each authorized researcher. The token shows a numeric code reinitialized every 60 seconds. Such a technique guarantees that the user of the remote connection is indeed the authorized researcher.

35. The research file can be processed only with the software provided and it is not possible for the researcher to download or print the data or results in his or her work zone.

36. When the researcher wishes to recover the results of the analyses, he or she must submit the request through the Internet portal. The results are examined by ISQ specialists in order to control the risk of disclosure (tables), in accordance with the ISQ's guidelines on the subject. If necessary, the tables are masked. Then, when they are in compliance, they are routed to the researchers.

37. Lastly, the user's work session is recorded and can be studied at any time (in real or delayed 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.

C. Added Value of the Service Platform

38. The service platform is an innovation in Québec. It meets an obvious need to place files resulting from the linkage of administrative or survey files at the disposal of researchers. In short, the new service facilitates research, particularly in the social and health fields. Furthermore, it provides other benefits, both for researchers and for stakeholders (authorities responsible for ensuring that privacy is respected and agencies holding administrative files).

39. Researchers have new means at their disposal:

- A researcher has access to an Internet portal for specialized services from which he or she can consult a dictionary that documents the variables available, use a search engine and follow up on his or her research request.
- The ISQ accompanies the researcher and provides help in preparing the request and in processing the research file.

- The ISQ performs for the researcher the operations of linking files without a unique identifier, according to rigorous methods that ensure the quality of the result.
- When the system is fully operational, the time required for authorization and obtaining research files should be reduced through a normalized process recognized by all stakeholders.
- The possibility for the researcher of having remote access to research files is, without a doubt, a substantial benefit, since it is not necessary for the researcher to travel.

40. Improvements in the current process for partners:

- An undeniable advantage, both for the Commission d'accès à l'information and for the holders of files, is that they should receive complete, clearer and better structured requests, since the ISQ conducts an initial analysis of the researcher's request that deals in particular with the availability and quality of the data requested, the technical feasibility of proceeding with the linkages required and the impact on the quality of the linkages and the masking to be done.
- The Commission should also see a reduction in the number of requests for changes that it receives following the research authorizations it issues.
- The location of the service platform at the ISQ provides guarantees that the confidentiality and security of the data will be ensured, given the legal oversight (statistical secrecy) and the normative framework of the ISQ in that regard.
- The Commission can be informed at all times of the stage the research has reached.

D. Current Status of the Project

41. The service platform is set up in two phases. During the first phase, the project design and the testing of components of the solution on the basis of a concrete research protocol are developed simultaneously. The second phase consists in the building and implementation of the service platform.

42. The service platform was set up in two phases. The first phase implied the design of the project and the testing of the components of the solution on the basis of a concrete research protocol. The second phase consisted in the building of the service platform itself. Besides experimenting with various functionalities of the overall solution (dealing with legal, personal information protection, technological, and file linkage and masking aspects, as well as with the organization of work in the partnership), the first-phase testing was used to study and compare the effectiveness of two linkage methods, namely, probabilistic linkage on the basis of direct identifiers and deterministic linkage on the basis of indirect identifiers.

43. The design of the service platform is now completed, and testing, on the basis of a research protocol called "The use of specialized services provided by the education system to children born prematurely" will continue until the spring of 2007. The testing, which began in February 2004, is not only to evaluate the components of the solution provided when the service platform was designed (dealing with legal, personal information protection, technological, and file linkage and masking aspects, as well as with the organization of work in the partnership), but also to test alternative approaches, especially to the linkage and masking of data.

44. The service platform should be in operation in the summer of 2007.

E. Financing and Continuity of the Project

45. The funding of the project is already secured for the first two phases of the establishment of the service platform by a Québec research support agency. The three years that follow will be a breaking-in period and will provide an opportunity to promote the new service and broaden the clientele to other public network researchers. The platform will be funded during these three years by Québec public agencies and research agencies.

46. In subsequent years, the bulk of the costs of operating the platform will be covered by the rates charged to researchers who wish to use the service. A market study conducted in February 2005 showed great interest in the service on the part of the research community, and measured the magnitude of potential requests for access.

It is likely that use will gradually increase over the years. The three-year breaking-in period, after implementation in the summer of 2007, will be used to promote the service. If the anticipated use materializes, it is expected that, by 2011, the platform should entirely pay for itself through research requests.
