ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Cigrang, Marc; Coder, John

Working Paper

Balancing Data Access and Data Protection: The Luxembourg Income Study Experience

LIS Working Paper Series, No. 57

Provided in Cooperation with:
Luxembourg Income Study (LIS)

Suggested Citation: Cigrang, Marc; Coder, John (1990) : Balancing Data Access and Data Protection: The Luxembourg Income Study Experience, LIS Working Paper Series, No. 57, Luxembourg Income Study (LIS), Luxembourg

This Version is available at:
https://hdl.handle.net/10419/160729

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Luxembourg Income Study Working Paper Series

Working Paper No. 57

Balancing Data Access and Data Protection: The Luxembourg Income Study Experience

Marc Cigrang and John Coder

August 1990

(scanned copy)

Luxembourg Income Study (LIS), asbl

BALANCING DATA ACCESS AND DATA PROTECTION:

THE LUXEMBOURG INCOME STUDY EXPERIENCE

by

Marc Cigrang
HAL and Technical Advisor, LIS

and

Lee Rainwater
Harvard University and Reseach Director, LIS

This paper was prepared for presentation a the Annual Meetings of the American Statistical Association held in Anaheim, California, August 6-9, 1990.

BALANCING DATA ACCESS AND DATA PROTECTION:

THE LUXEMBOURG INCOME STUDY EXPERIENCE

INTRODUCTION

The Luxembourg Income Study(LIS) experiment began in 1983 under the joint sponsorship of the government of Luxembourg and the Center for Population, Poverty, and Policy Studies, a nonprofit research group also located in Luxembourg. The primary objective of this experiment was to promote research on the distribution of income, the level and characteristics of poverty populations, and the general economic situtation of households and families in an international context.

Since its beginning in 1983, the experiment has matured into a cooperative research project with a membership that includes countries in both Western and Eastern Europe, North America, and Australia. At this moment, there are a total of 14 member countries(see Appendix A). Microdata from household surveys conducted by the member countries have been donated to the project and reside in databases maintained at the Centre Informatique de l'Etat, the computing center for the government of Luxembourg. This microdata consists of social, demographic, economic, labor market, and geographic information about the households and families participating in these surveys.. Researchers are not allowed direct access to the data as some of the participating countries have imposed restrictions on their use even though all data are anonomized to remove any names, addresses, or other identifiers. Access is provided using the SPSSX statistical package and the LIS user interface developed in Luxembourg. The system is designed mainly to provide remote access to users with connections to the EARN/BITNET computer mail network. Use of the data is restricted to academic and policy analysis research. Use for commercial purposes is prohibited.

The purpose of this paper is to describe the development and current status of the LIS project with emphasis on the systems that manage and monitor access to the database. This system represents a balance between the main goal of the project, that of promoting international comparisons of economic well-being, and obligations to those countries that have supplied data under agreement that the data be protected.

LIS PROJECT DEVELOPMENT

In the early stages of the study traditional solutions were sought for the problem of how to obtain and manage the data needed to examine differences in economic well-being. There were four basic elements to this potential solution. The first element would be to identify the issues that should be studied through consultation with reseachers interested in international comparisons in these areas. The next element would be development of specifications for tabulations that provide the kind of information needed to investigate these issues. The third element would be locating resouces in the statistical agencies or other research groups within each country that could create these tabulations. Finally, the fourth element would be to develop methods for disseminating and documenting these tabulations.

At some point during initial planning stages an untested alternative to this traditional approach emerged. This new solution called for acquisition of the survey mircrodata and construction of a database maintained in a central location. The existance of the microdata would allow researchers to analyze the data without the restriction imposed by previously tabulated data which is often inappropriate for their specific needs.

This model presented some very large problems. First, while many countries allow public access to microdata files derived from surveys conducted by their national statistical agencies, many others do not. How could those countries who do not permit public access be persuaded to participate by donating microdata? A second, but closedly related question, was how could access to the data be provided, both from technical and data protection perspectives? Third, even if countries are willing to provide microdata, who could supply the computing resources necessary to maintain the database and service the requests of researchers?

Satisfactory answers to some of key these questions were found largely through the cooperative efforts of the Center for Population, Poverty, and Policy Studies(CEPS), the government of Luxembourg, and the Centre Informatique de l'Etat(the computing center for the Luxembourg government). The necessary computing resouces were provided by the Centre Informatique(CIE). The CEPS became the parent organization to which the LIS is attached. It acted to provide a base of operations, allocated office space, acted as liasion with the government of Luxembourg and the CIE, and provided funding in the initial stages of project development.

Once these major issues regarding resources had been resolved, detailed work began in three areas. First, organizational aspects of the project needed to be formalized and countries needed to be contacted regarding participation. Second, a system needed to be developed that balanced access to the data with the data security required by particpants in the project. Third, sources of future funding needed to be found. The remainder of this paper covers the details regarding the second area of work, that of developing the LIS access system.

THE LIS ACCESS SYSTEM

The LIS access system was developed with two basic principles in mind. First, in order to promote international research and extend use of the data to a large number of individuals a system was needed that permitted users to access the data from remote sites. Second, the data needed to be "protected" since many member countries could be expected to supplied the survey microdata only under the condition that it be solely maintained at the CIE in Luxembourg. Other factors that determined the final design of the system included 1) access to widely available statistical procedures, 2) "turnaround" time for user requests and the ability to handle large volumes of work, and 3) limitations on the system disk storage.

SYSTEM OVERVIEW

The LIS operating system is basically a batch processing system in which remote users submit job requests in the SPSS-X statistical package language utilizing linkages with the EARN/BITNET computer network. For example, a researcher in Australia can develop an SPSS-X program that specifies the variable names used by the LIS system and send that program to Luxembourg via the EARN/BITNET network. This request is processed and the resulting statistical output is returned to the researcher also via EARN/BITNET.

If desired the system is capable of operating in automated mode for 24 hours a day. A system "sleep" state is maintained until a job request is received. At that time the system is activated. Since the system services users around the world, this feature helps provide faster turnaround of jobs and helps distribute the daily workload.

SYSTEM COMPONENTS

Quite naturally the design of the LIS processing system is tied directly to the environment. In this case the environment is an IBM 4381 mainframe operating under VM/CMS. Components of the system include a mailbox or communications package, a work distribution and monitoring package, a batch processing package, a database component, and access by the batch the processing area to the SPSS-X statistical package. Since the VM/CMS operating system uses the concept of "virtual" machines, this concept became an integral part of the access and protection mechanisms.

MAILBOX. The mailbox package is both the brain and heart of the LIS system. It resides on a virtual machine and is responsible for a number of vital duties. All requests for LIS services must be received via the mail machine. As the mailbox also receives a variety of other messages, etc. it is designed to receive and distinquish between messages and job requests in an automated manner, i.e. without human intervention. Once text has been identified as a job request, the mailbox then screens

the request to remove any extraneous "headers or "trailers" that may have been included by the user or inserted by system mailers that control the flow of mail in the EARN/BITNET network. Checks are made for a valid user identification code and password that must be preassigned through a process that requires each user to register. This registration process also establishes the EARN/BITNET mailing address to which the statistical output is to be sent. All of this information on each user is maintained in a database for access by the mailbox and the LIS staff.

Following the check for identification and password the SPSS-X request is reviewed for content. This content review attempts to identify procedures, etc. that could result in listing or copying of data for individual cases. Submissions that fail this review are sent for detailed examination by the LIS staff. Those completeing this review process successfully are sent to the work distribution center for further action.

After completion of the SPSS-X execution the mailbox receives the output listing from the work distribution center and conducts some additonal reviews of the contents. Output failing these checks are sent to the LIS staff for further review. If no violations are found the output is returned to the user via EARN/BITNET.

WORK DISTRIBUTION AND MONITORING(WDM). The work distribution and monitoring functions also reside on individual virtual machines. The WDM machine serves three functions. It is the link between the mailbox and the batch functions. It also distributes, manages, and monitors the work flow, and it provides security for the data. All communications between the mailbox and the batch area of the LIS system is conducted through the WDM. The WDM will not accept communications from any other source. It receives job submissions from the mailbox and distributes them to the batch area where executions take place. It receives output from the batch area and sends them it the mailbox where it is reviewed and returned to the requestor. The WDM also allows the LIS staff to monitor the processing system by displaying job statuses and queue information.

BATCH PROCESSING. Execution of the job submissions take place exclusively in the VM machines dedicated for batch processing. These are the only machines in the system that have access to the relational database that contains the microdata. Upon receipt of a job request from the work distribution area the batch machine "selects" only variables needed for the request, creates an SPSS-X system file, and executes the request. Upon completion the statistical output is sent to the work distribution center.

DATABASE. The survey microdata for each country is stored in IBM's relational database system. A relational database was chosen to store the LIS data because it allowed an efficient use of limited computer disk storage and because it provided some additional data security. A relational structure was chosen because the survey data available to LIS included both household and person(household member) level information. As the number of household members can differ widely, use of the relational structure avoided allocation of storage space for some fixed, maximum number of household members to each household. In addition, access to the database is governed by the "creator", a fact which allows use of the data to be tightly controlled.

ELEMENTS OF DATA PROTECTION

While the basic elements of data protection provided by the LIS system were mentioned in the previous discussion in relation to the system components these elements should be addressed specifically. The data protection scheme devised for LIS has two basic objectives. First, it is designed to prevent copying of the microdata on a case by case basis, that is, transfer of the microdata from the database to some other location. Second, it is designed to prevent the linkage of the microdata in the database with the names, addresses, or other information that would identify a specific individual in the population. To achieve these requirements we have introduced the following procedures:

a) anonomize all microdata records to remove all identifiers such as name, address, questionnaire number, etc.

b) round all income figures to remove "exact" amounts that, in some cases, may have been taken from administrative record sources.

c) require that all users of the data register with the LIS staff to obtain a user identification and a password.

d) require that all output from job requests be sent only to the EARN/BITNET address specified by the user and registered in the user database.

e) allow access only through the LIS interface using the SPSS-X statistical package.

f) perform strict reviews of the job requests and statistical output.

g) use of system of independent virtual machines that separates users from the microdata.

h) storage of the microdata within a database that can be accessed only by those individuals who have been granted proper authority.

Accessibility to individual databases and the privelege to access the system are governed by information contained in two other relational database tables independent of the microdata. If problems are detected with a particular database are encountered access can be temporarily suspended by changing access codes in the database while use of the data for other countries continues. Access to the system can be temporarily or permenantly terminated at the user level in a similar fashion.

DATABASE CONTENT AND OTHER ASPECTS OF LIS

The LIS database contains a wide variety of information on the social structure and economic status of households in the particpating countries. As the data contained in the LIS system were derived from existing survey operations, the kind of information and level of detail available for each individual country database differ. In order to preserve these differences and yet provide some common structure to the database a compromise was reached. A list of standard demographic, social, labor market, income, and geographic variables was established (see Appendix B). Most of these are very common in surveys of household economic situations. For example, number of houshold members, age and sex of household head, number of children, occupation and industry of work, hours worked, earned income, retirement income, etc. Each data item for a particular survey was then assigned to one of these standard categories. In most cases, especially for categorical variables such as occupation or education, the content of the variable differs from country to country because the social institutions and the country classification schemes used in the surveys differ.

The variable names and value label that identify the meaning of numerical codes for categorical variables are maintained in relational databases independent of the microdata. These descriptor databases are accessed to provide this information in formation of the SPSS-X system file.

Interactive software is also available that allows onsite users to browse though descriptions of the database contents without having access to the data itself. These content databases are also the source for creation of online documenation packages that can be sent to users via the EARN/BITNET network.
Other onsite services include a free-text database containing special notes and other information about specific variables and an automated job setup package. Users can add notes to or make queries of this free-text area to find comments by other users or the LIS staff concerning previous findings, data problems, etc. Users unfamiliar with the LIS job submission specifications can use the automated job setup package to choose country databases and variables to build the skeleton for an SPSS-X job request.

Work is underway this summer to begin building an additional database containing information about the tax and transfer programs for each country. This will improve the ability of users to contrast differences between countries in the structure of social programs with respect to eligibility criterion, payment levels, etc. It will also aid in examining differences regarding income, payroll, and other types of taxes. This tool is likely to be developed for use on personal computers with copies made available on request.

CONCLUDING REMARKS

During the past seven years the Luxembourg Income Study has passed from the experimental stage to a well-recognized cooperative international research project in the fields of sociology and economics. It provides easy, remote access to researchers throughout the world, balancing this access with the data protection required by the project's members. There are at this point about 150 registered users. The number of job requests has grown steadily and now averages nearly 50 per day. The number of country databases has grown steadily as well reaching a total of 22 by the Spring of 1990. New country databases are expected later this year. These include data for Hungary, Czechoslovakia, and East Germany.

The growth of LIS presents a number of problems. Managing the growing size of the database and the growing number of job requests is a difficult task. The survey sample sizes for each country contain information for an average of about 9,000 households and more than 20,000 household members. The computer resources needed store the data and execute the job requests is substantial. Innovative solutions to these problems are being persued and will be part of the design of the next generation of the LIS system.

Inquires concerning membership or access to the Luxembourg Income Study can be made by sending a message to the LIS office in Luxembourg at the address listed below or by BITNET at SSLISBB@LUXCEP11.

LIS
Boite Postale 65
L-7201 Walferdange
Grand-Duche de Luxembourg

APPENDIX A

SUMMARY OF CONTENTS OF THE LUXEMBOURG INCOME STUDY DATABASE

FIRST TIME PERIOD: 1979-81

COUNTRY	REFERENCE YEAR
Australia	1981
Canada	1981
France	1979
Germany	1981
Israel	1979
Netherlands	1983
Norway	1979
Sweden	1981
Switerland	1982
United Kingdom	1979
United States	1979

SECOND TIME PERIOD: 1984-87

COUNTRTY	REFERENCE YEAR
Australia	1985
Canada	1987
Germany	1984
Italy	1986
Luxembourg	1985
Poland	1986
United States	1986

APPENDIX B

SUMMARY OF LIS DATABASE CONTENTS

Household Variables

1) Number of household members
2) Number of children under age 18
3) Age of youngest child
4) Number of earners
5) Geographic region
6) Household composition
7) Tenure
8) Total household earnings (sum of all household members)
9) Total amount of income received by type of social transfer
10) Total amount of income received by type of private transfers
11) Total amount of income from capital(property income)
12) Total amount of income and payroll taxes
13) Total amount of employer contributions for social insurance
14) Total amount of employee contributions for social insurance

Person Variables

1) Age
2) Sex
3) Relationship to household head
4) Nationality or Ethnicity
5) Marital Status
6) Educational level
7) Occupational training
8) Occupation of work
9) Industry of work
10) Type of worker(socio-occupational status)
11) Labor force status
12) Weeks worked full-time and part-time during the year
13) Hours worked per week
14) Disability status
15) Wage and salary income
16) Social retirement income
18) Private and public pension income
19) Unemployment compensation payments
20) Income and payroll taxes